Towards Sentinel 2 based environmental contamination monitoring

Christian Köhler

Institute for Mine Surveying and Geodesy, TU Bergakademie Freiberg

ABSTRACT:

Supporting environmental monitoring with remote sensing data considerably increases cost effectiveness and reliability of traditional, manual solutions. Additionally, by means of automation, it bears the potential to prevent contaminations or disasters through the availability of timely and spatially dense data. To this end, we investigate the possibility to monitor gas and oil pipelines of a storage cavern by using optical, multi spectral data from the Copernicus Sentinel 2 satellites. Due to a lack of known disasters/contaminations, we resort to a monitoring approach based on statistical outliers. First results demonstrate the general capability of our approach to detect contaminations and generate warnings.

1 Introduction

Mining activities strongly influence the environment, e.g. by causing ground movements or producing contaminated waste dumps and water entities during the active operation and rehabilitation phases. Thus, monitoring and managing the contamination is essential as it reduces environmental impact and therefore increases social acceptance.

Terrestrial contamination monitoring in terms of manually taking ground samples is not cost effectively implementable and lacks spatial and temporal resolution. Providing the required spatial and nearly time continuous resolution, contamination monitoring with remote sensing data based on the Copernicus Sentinel missions 1 and 2 is a highly promising approach towards cost effective monitoring systems. However, contamination, e.g. low pH values due to acidification or high concentration of deleterious elements in the ground or waste dumps, is not directly measurable with remote sensors. New approaches are needed to derive proxies for contamination monitoring. These have to be calibrated on direct measurements from available data.

For this purpose, aim at an effective contamination monitoring based on multi-scale data by undermining statements inferred from space-borne Copernicus Sentinel 1 and 2 data with aerial imagery and in-situ ground measurements. Here, we focus on the remote sensing part based on the optical multi-spectral data from Copernicus Sentinel 2 satellites by means of investigating a case study: Contamination monitoring along media pipelines for oil and gas storage caverns.

The repeated filling and drainage of oil and gas storage caverns results in underground movements that can exceed several meters. The induced stress - and possible failure – of the present media infrastructure is a major threat to the health of environment and people. An early warning system due to timely dense monitoring of contaminations is a major contribution to operational safety.

An overview over the selected investigation site is given in Abb. 2 (a), showing the region in false colours, overlaid with the course of the pipeline.

2 Monitoring Approach

Our goal is to generate contamination warnings from remote sensing data.

For this investigation, the general idea is to use the optical, multi-spectral (MS) data delivered by Copernicus Sentinel 2 satellites. Of course, using any other suited source of optical MS data, in particular from UAV, is possible, but not subject of the current considerations. After data preprocessing, MS data are available as multi-layer Geo-tiff files, where the values in each layer correspond to the surface reflectance of the multi-spectral bands.

The MS reflectance data serve as input for the analysis, which is performed within our defined region of interest. The analysis is based on image objects (IO, units of neighbouring pixels with similar spectral properties) and performed in a spatial and spatio-temporal context.

As a result, binary warning codes are extracted for each image object, encoding contamination being present or not present in the corresponding IO.

In the following, we detail on the processing steps from data acquisition via image pre-processing to extraction of contamination warnings.

2.1 Data preparation

2.1.1 Data sources

In general, different multi-spectral data sources can serve as basis for our upcoming investigation and analysis of environmental contaminations. Suitable MS data are characterised by the following conditions: First, data are available as images and cover the region of interest. Second, the pixel size should allow the localisation of possible contaminations. Third, at least a few MS bands in the visible to NIR range are available in order to calculate spectral indices.

MS Data from satellites such as Sentinel 2, Landsat 7 and Landsat 8 are suitable, but also hyperspectral cameras mounted on UAVs or planes fulfil the requirements.

As outlined above, we use multispectral data from Sentinel 2A and 2B satellites. These are free of costs and publicly available at https://scihub.copernicus.eu/dhus/. Further, they have high spatial, spectral and temporal resolution (see Tab. 1) and the level 2a data products readily provide surface reflectances (ESA 1), (ESA 2).

Tab. 1: Multi-spectral bands of Sentinel 2A.

Band #	Central Wavelength [nm]	Bandwidth [nm]	Description	Spatial Resolution [m]
1	442.7	27		60
2	492.4	98	Blue	10
3	559.8	45	Green	10
4	664.6	38	Red	10
5	704.1	19	VRE	20
6	740.5	18		20
7	782.8	28		20
8	832.8	145	NIR	10
8a	864.7	33	NIR	20
9	945.1	26		60
(10)	1373.5	75	Cirrus	60
11	1613.7	143	SWIR	20
12	2202.4	242	SWIR	20

We used scenes of Sentinel 2A and 2B covering the use case site from the year 2018.

2.1.2 Pre-Processing

In a first pre-processing step, we prepare the data for further use by generating a common manageable data format.

We start with stacking the optical bands, the bands encoding the cloud probability and land cover classification into one multi-layer Geo-Tiff. Further, all pixels are resampled to the highest spatial resolution, i.e. 10m for Sentinel 2. Finally, we extract a subset of the scene, covering our region of interest in order to reduce memory consumption and processing time.

For Sentinel 2 scenes before 2018, the level 2a products are not readily available. Then, an additional step, generating surface reflectances from the level 1c product with Sen2Cor (Sen2Cor), is necessary.

The above pre-processing is implemented in python, using python libraries (refs).

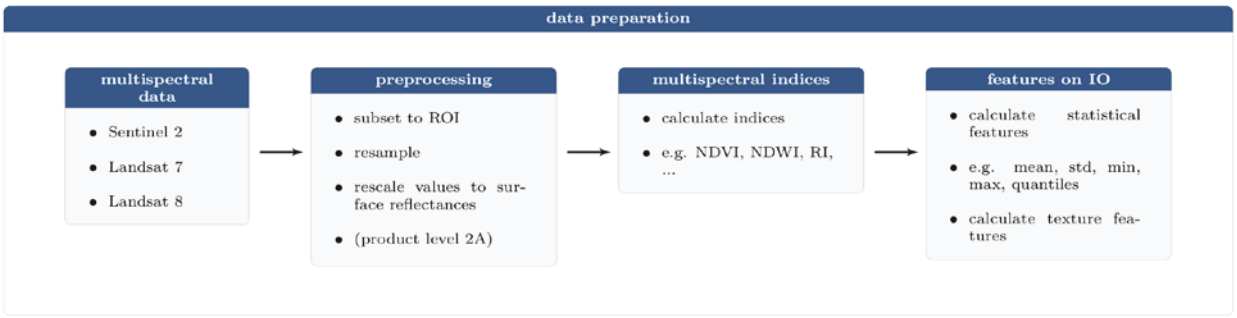


Abb. 1: Data pre-processing steps.

2.1.3 Multi-Spectral Indices

Multi-spectral indices are new variables that are calculated per pixel from the multi-spectral bands. In several applications, the use of these indices turned out very useful, as these are designed such, that they emphasize certain properties of the recorded scenes which are not prominent in the original data. E.g., the very well known Normalized Difference Vegetation Index (NDVI) is calculated as

$$NDVI = \frac{R_{NIR} - R_{RED}}{R_{NIR} + R_{RED}},$$

where R_{RED} and R_{NIR} denote the reflectances of the red and near infrared band (i.e. band 4 and 8 for Sentinel 2), respectively. It allows to distinguish between vegetated ($NDVI > 0.3$) and not vegetated ($NDVI < 0.3$) areas (see Abb. 2 (b)).

For our upcoming investigations, we selected a large number of vegetation, soil and water radiometric indices (see Tab. 2). This selection is not comprehensive and will probably be even further reduced, once the most meaningful indices are identified.

Tab. 2: Vegetation, soil and water radiometric indices used for further analysis. The often similar equations and the use of mainly the red, green and NIR bands render information redundancies probable.

Name	Long Name
Vegetation radiometric indices	
SAVI	Soil Adjusted Vegetaion Index
MSAVI2	second Modified Soil Adjusted Vegetation Index
DVI	Difference Vegetation Index
RVI	Ratio Vegetation Index
GNDVI	Green Normalized Difference Vegetation Index
GSWIR	Green SWIR
ARVI2	
NDI45	Normalized Difference Index
NDVI	Normalized Difference Vegetation Index
MTCI	Meris Terrestrial Chlorophyll Index
MCARI	Modified Chlorophyll Absorption Ratio Index
S2REP	Sentinel-2 Red-Edge Position Index
IRECI	Inverted Red-Edge Chlorophyll Index
PSSRA	Pigment Specific Simple Ratio Algorithm
Water radiometric indices	
NDWI	Normalized Difference Water Index
NDWI2	Second Normalized Difference Water Index
NDPI	Normalized Difference Pond Index
NDTI	Normalized Difference Turbidity Index
Soil radiometric indices	
BI	Brightness Index
BI2	second Brightness Index
RI	Redness Index
CI	Colour Index
STR	
WETNESS	

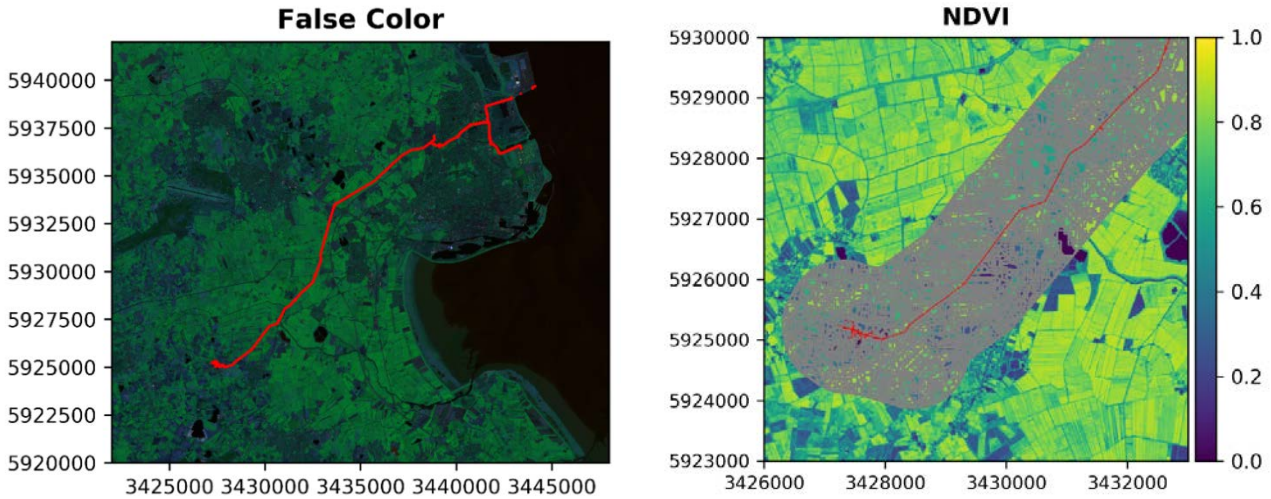
2.1.4 Features on image objects

Up to this data pre-processing step, we remained within the traditional pixel based view. That is, all the pixels have the same size, shape and are isolated from each other.

However, for image analysis it is useful to group pixels together into image objects (IO). The grouping is performed by applying the Mean Shift algorithm and using all the available spectral bands as input data. The resulting vector shapes (polygons) then embrace pixels with similar spectral signatures. The size of the IO and the allowed intra-object spectral difference enter the algorithm as parameters.

The IOs allow the extraction of additional features for each IO compared to the pixel based view. For our analysis we use statistical features such as minimum, maximum, mean, standard deviation and quantiles of underlying data bands.

As a result of the data preparation step, we have vector shapes (the image objects) attributed with numerous features. These are the min, max, mean, standard deviation and the x-quantile (x ranging from 0.1 to 0.9 in steps of 0.1) values for each of the indices listed in Tab. 2.



(a) (b)

Abb. 2: (a) False color composition of the Sentinel 2 scene of the region of interest using the bands (r,g,b)=(4,8,11). The location of the pipeline is shown in red. (b) A detail of the scene in (a). The color corresponds NDVI values. Overlaid are the image objects used for the analysis, colored according to the mean value of NDVI within the IO.

2.2 Data analysis

In the following, we want to present the steps towards a remote sensing based supported contamination monitoring of the gas and oil storage cavern.

The potential contaminations are oil and brine, leaking from one of the pipelines.

In the vicinity of the pipelines, we deal mainly with vegetation covered soil. This prevents a direct detection of the contaminants and we resort to indirect detection by recording the effects of the con-

taminants on plant health. Plant health in encoded in different vegetation radiometric indices by low or high values compared to healthy plants.

In addition, we meet one more difficulty: Among our available data we have only one verified example of an area contaminated with brine. This disqualifies approaches using supervised classification or training models, as the number of positive (=contaminated) training samples is too small and we do not have examples of different contaminations (i.e. oil). Moreover, all developed algorithms that depend on training, be it within established methods (e.g. SVM etc.) or by hand (e.g. thresholding indices/combinations of certain indices, decision trees, etc.) are prone to be overfitted to this specific example of contamination.

To resolve this issue, we base our analysis on statistical outlier detection. This approach assumes, that the contaminations are clearly distinct from normal behaviour (in terms of e.g. index values) and characterised by extreme values that can be found close to the margins of the underlying statistical distribution.

2.2.1 Outlier detection

The numerous attributes we calculated for the image objects can be viewed as random variables. Especially, mean values and standard deviations can be calculated by considering many image objects at once (e.g. all IO within a certain radius). A specific chosen IO then can be compared to these inter-IO-characteristics in terms of the distance between the mean value and the actual IO value. This distance can be defined in different ways.

For univariate random variables, the z-value Z measures the distance of a random variable X to its expectation value in units of its standard deviation: $Z = (X - \mu)/\sigma$. So unusual, i.e. extreme values would translate into high Z. E.g. for normally distributed values, less than 0.3% have z-values larger than 3. By introducing a threshold, IO outliers are identified as the ones with z-values exceeding it.

Similarly, the interquartile range $IQR = Q_3 - Q_1$, with Q_1 and $Q_3$the first and third quartile, allows outlier detection by defining all observations below $Q_1 - 1.5\,IQR$ or above $Q_3 + 1.5\,IQR$ as outliers.

A multivariate version of the z-value is the Mahalanobis distance (Mahalanobis, 1936). It is defined as $D(x, y) = \sqrt{((x - y)^T \sum^{-1}(x - y))}$, where x, y are two realisations of a multivariate random variable and $\sum$ their covariance matrix. It measures a distance between two multidimensional points, where each dimension is scaled by its standard deviation. Therefore, the Mahalanobis distance is mapping values of numerous IO attributes to one single value.

A second group of methods for outlier detection is clustering algorithms. Clustering algorithms are suited for multivariate data and try to divide all available data into one or more groups of similar characteristics.

The single class support vector machine expects one parameter, the fraction of data that are outliers and accordingly determines a mutli-dimensional region. Outliers are points outside that region.

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm builds several clusters based on density. It is capable of disregarding noisy observations by not assigning them to a cluster. These are our outliers.

Both clustering algorithms depend on the choice of their parameters. Thus, careful choice is mandatory in order not to overfit towards the given specific examples of contamination.

3 Preliminary results

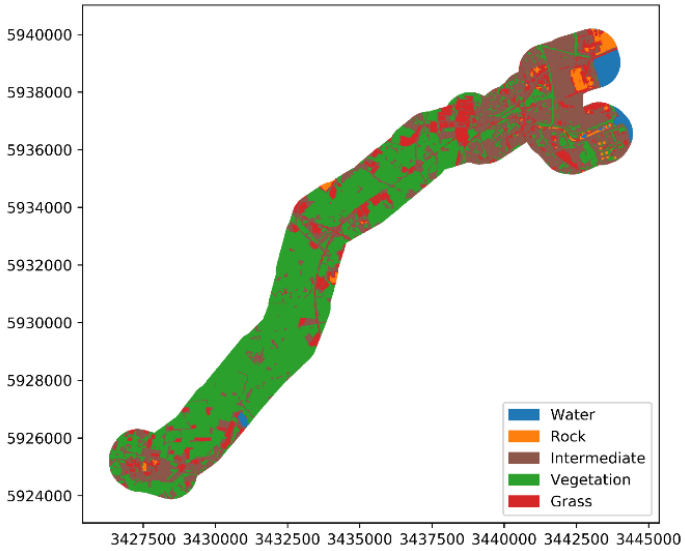


Abb. 3: Course of the pipeline corridor and land cover types

We performed our detection of contaminations only within a corridor of 2km width around the course of the pipeline. Further, for identifying outliers, we restricted comparisons of IO observables to IO with the same land cover type. The pipeline corridor and the land cover types of water, rock, intermediate, grassland and vegetated soil is shown in Abb. 3. There, the area of IO is colour coded according to their land cover type for the IO within the corridor.

In Abb. 4 the results for one possible realisation of generating a contamination warning based on Mahalanobis distances is shown. This specific realisation is based on the knowledge of one contamination event, where brine was spilt onto vegetated soil at the location marked as blue dot in the lower left corners of the scenes. We hand picked five attributes (standard deviation of NDVI, NDWI, NDTI, RI, CI) by comparing the z-values of the IO close to the contamination site from before and after the incident. If the z-value of a certain attribute was below a threshold of 3 before and above 3 after the contamination, it was selected.

Generating warnings from thresholding the z-values alone lead to a large number of false positive alarms (not shown). Therefore, the Mahalanobis distance (MD) was calculated for the combination of these five attributes, depicted in Abb. 4, top row. Introducing a threshold of 4 for the MD, identified a reasonably small number of IOs as contaminated (marked as red patches in Abb. 4). By taking into account only the contaminated IOs that are connected to the pipeline, we identify roughly one false alarm in the pre contamination scene (Abb. 4 top left) and roughly two false alarms in the

post contamination scene (Abb. 4 top right). More importantly, the true contamination was detected, as indicated by the red patch (=contamination warning) at the contamination site (blue dot).

Repeating this procedure, but deriving the MD using the mean values of the five indices instead of their standard deviations, leads to the situation displayed in Abb. 4 bottom row. While having almost no positive false alarm for the pre-contamination scene (bottom left), a large number of false alarms appears for the post-contamination scene. Even worse, the actual contamination is missed.

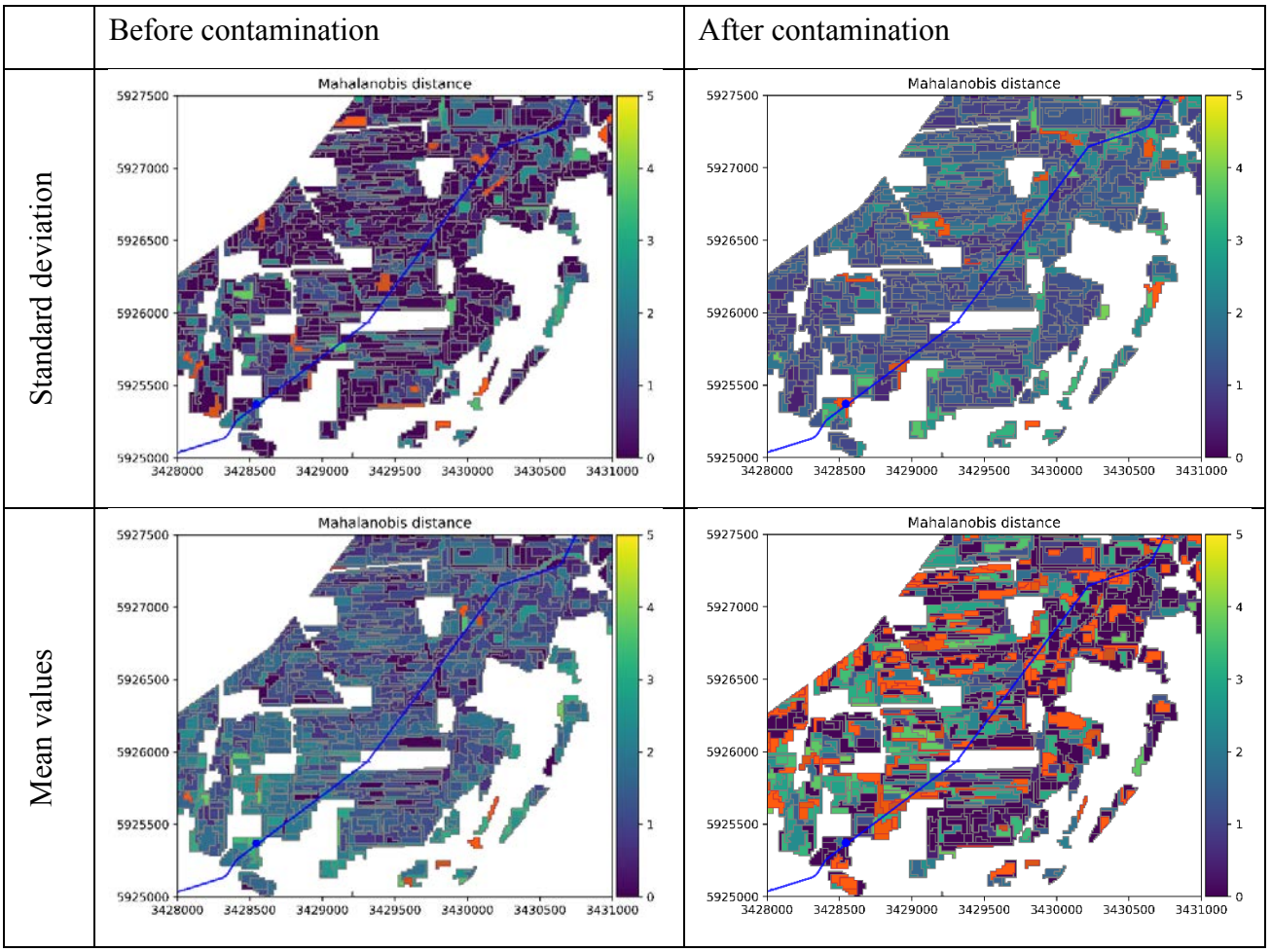


Abb. 4: Image objects of the vegetated land cover type colour coded with Mahalanobis distance and overlaid with generated warnings in red. The Mahalanobis distance is calculated for standard deviations (top row) and mean values (bottom row) of the NDVI, NDWI, NDTI, RI and CI indices. Scenes are selected from before (left column) and after (right column) a contamination with brine occurred at the site marked as blue dot (lower left corner of scenes).

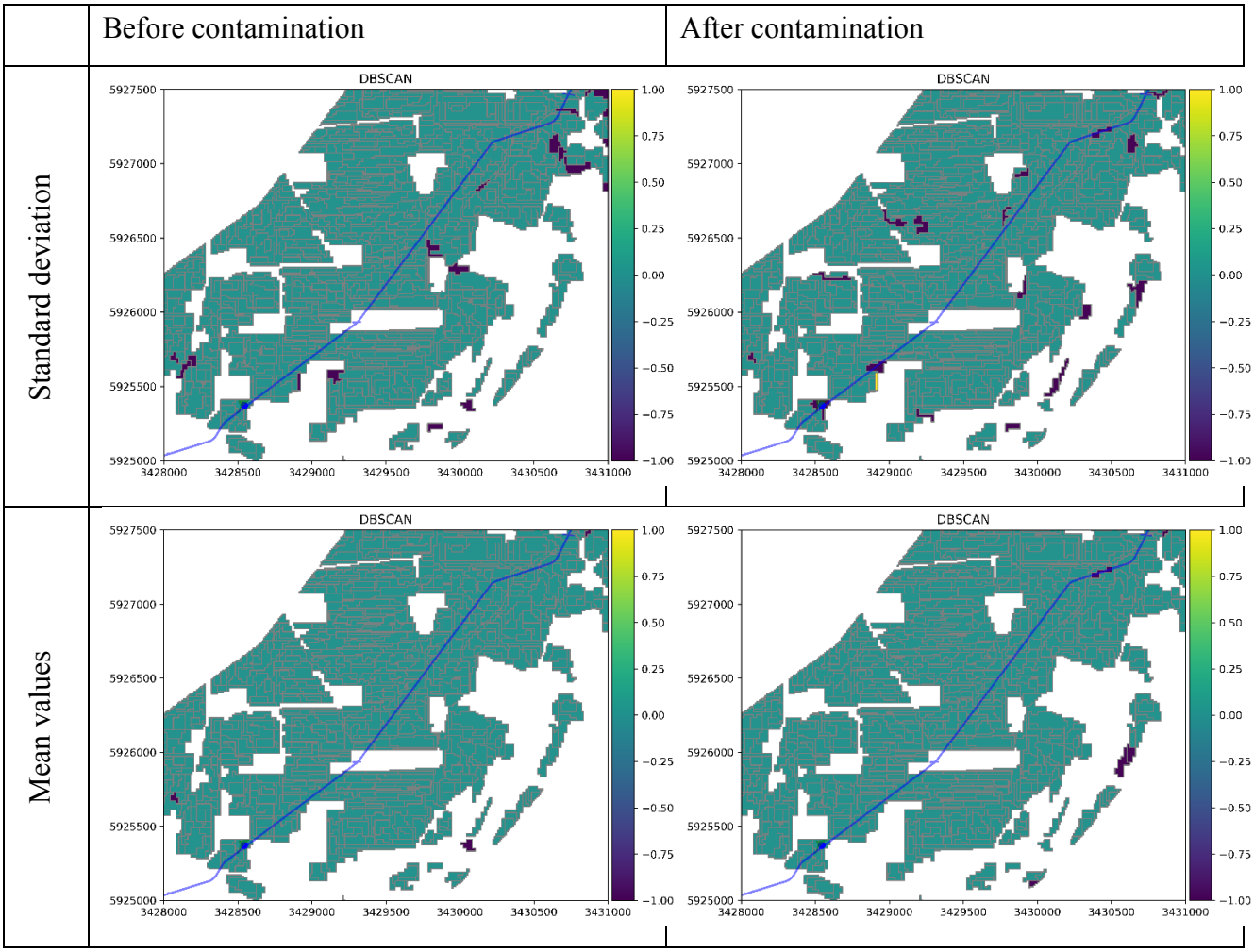


Abb. 5: Image objects of the vegetated land cover type color coded with the identification number of clusters extracted by DBSCAN (eps=0.6, min samples 6). Attributes are again standard deviations (top row) and mean values (bottom row) of the NDVI, NDWI, NDTI, RI and CI indices. Scenes are selected from before (left column) and after (right column) a contamination with brine occurred at the site marked as blue dot (lower left corner of scenes). Outliers are painted dark.

In Abb. 5 we used the DBSCAN clustering algorithm with hand picked parameters eps=0.6 and minimum samples=6. For both attribute sets of mean values and standard deviations of (NDVI, NDWI, NDTI, RI, CI), the number of IO that are directly on the pipeline and show false alarms is much smaller compared to the MD approach. Unfortunately, the case using mean values again misses to identify the present contamination. In contrast, using standard deviations detects the contamination, but also produces additional false alarms.

4 Summary and Outlook

Altogether, we showed, that indirectly detecting contaminations with optical remote sensing data from Sentinel 2 satellites is possible. Even the rather simple approaches of thresholding the MD or clustering with DBSCAN using five selected attributes proves successful. However, the strong dependence on the selected attributes is emphasized and the danger of an overfitted solution is real.

Nevertheless, the failure of contamination identification when using mean values teaches, that absolute values of indices/attributes might not be the best choice. On the other hand, successfully using attributes based on standard deviations of indices suggest, that contaminations are characterized by an increased spread of index values within the IO. Therefore, the use of e.g. texture based attributes seems promising.

Further, we did not even came close to scan the vast area of possible index and (not only) statistical properties combinations. Importantly, we observe, that the successful combination of indices consist of a vegetation, two water and two soil radiometric indices (at least for the one given contamination). That indicates, that even when considering vegetated areas, data from all land cover classes are useful. To this end, a purely data based approach is tempting. This could start with the twelve MS bands and follow the road of dimensionality reduction (PCA, MNF, …) , clustering and outlier detection. Lastly, a more sophisticated combination of several approaches, fuzzy control systems etc. deserve an attempt.

LITERATURVERZEICHNIS

ESA 1. Processing Levels. [Online] [Zitat vom: 11. April 2019.] https://earth.esa.int/web/sentinel/user-guides/sentinel-2-msi/processing-levels/level-2.

ESA 2. Resolutions. [Online] [Zitat vom: 11. April 2019.] https://earth.esa.int/web/sentinel/user-guides/sentinel-2-msi/resolutions/radiometric.

Mahalanobis, P. 1936. On the generalised distance in statistics. Proceedings of the National Institute of Science of India. 1936, Bd. 2, 1.

Sen2Cor. [Online] [Zitat vom: 11. April 2019.] http://step.esa.int/main/third-party-plugins-2/sen2cor/.